Mathematical Biology and Bioinformatics. 2012. V. 7. №1. P. t3-t15.
URL: http://www.matbio.org/2012/Pugacheva_7_t3.pdf
Translation from original Russian text © Pugacheva V.M., Korotkov A.E., Korotkov E.V., 2012 published in Matematicheskaya biologiya i bioinformatika, 2012, V. 7. №2. P. 461–475. URL:http://www.matbio.org/2012/Pugacheva_7_461.pdf

========== **BIOINFORMATICS** ==========

UDC: 577.212.2

Searching for pair points of triplet periodicity phase shifts in the genes of 17 bacterial genomes

©2012 Pugacheva V.M.[*1], Korotkov A.E., Korotkov E.V.[1]**

[1] *Bioengineering Centre of RAS, Russian Academy of Sciences, Moscow 117312, Russia*

Abstract. This paper presents a mathematical method developed for searching for pair phase shifts of triplet periodicity. Such phase shifts could be potential frameshifts in genes resulting from insertions of quite long DNA fragments. We developed software based on the proposed mathematical method and checked if there were pair phase shifts of triplet periodicity in genes of 17 bacteria genomes. The results show that in these 17 genomes about 1 percent of bacteria genes have pair phase shifts of triplet periodicity. This paper also describes a method developed for visualization of pair phase shifts of triplet periodicity and gives examples of such shifts. The research results were partially confirmed by the search for similarities in amino acid sequences that had been made by alternative reading frames. The connection between pair phase shifts of triplet periodicity and frameshift in genes is discussed.

Key words: *triplet periodicity, reading frame, frameshifts, phase.*

INTRODUCTION

Short insertions of DNA fragments can be rather common in genes [1, 2]. Unless the lengths of these inserts are multiple of three DNA bases, it leads to a frameshift after the insertion position. Such inserts can significantly change the amino acid sequence of the gene and it is important to understand their contribution to the frameshifts [3–5]. At present, there are two groups of mathematical methods that are used to search for frameshifts. Both of these groups require extra information in addition to the analyzed DNA sequences. The additional information for the first group of methods is databases of amino acid sequences and software for similarity search [6–9]. Amino acid sequences by alternative frames are generated by these algorithms and then are searched for similarities in the database. If the similarities are found, we can suppose that the analyzed gene has a frameshift. The second group of methods uses the nucleotide sequence of an analyzed gene to find frameshifts. A sample of genes in which an existing frameshift is already known is used for additional information [10–14]. As a result some general properties of the DNA areas where such shifts have been already found are searched for in an analyzed gene. Weight matrix, frequencies of *k*–tuples, HMM models, neural networks and other mathematical approaches are used as a way to describe such general properties [10–14].

These methods have some drawbacks that limit their application. These drawbacks are connected with the usage of additional data. The restrictions for the first group of methods are that amino acid sequence database must contain a sequence which may exist before the frameshift occurred. And it should have significant similarity with amino acid sequences

[*] virentis@gmail.com
[**] genekorotkov@gmail.com

created by the alternative frame. Frequently there is no such sequence in database or there is no significant similarity for it. Therefore search for the frameshifts by this method becomes impossible. Methods of the second group have the restriction of a different kind. These methods use the idea of creation general statistic properties of gene regions, which have a known frameshift. Using these properties some permissive mathematical rules are created to search for frameshifts in other genes. But, as it was shown previously [15], statistic properties of gene sequences may be different and therefore genes may have different classes of triplet periodicity. On combining different genes with known frameshifts many statistic characteristics of sequences may become mild and that can considerably reduce the power of frameshift search.

The previous approach for searching for triplet periodicity phase shift had been used to find potential frameshifts in genes [16, 17]. This approach has a significant difference from the methods that were mentioned above. It doesn't require extra information such as amino acid sequences database (like the methods of the first group do) or DNA base sequences with known frameshifts (like the methods of the second group do). To identify frameshifts we use the idea of gene triplet periodicity and triplet periodicity phase shift [15–17]. Triplet structure of DNA sequences encoding proteins is common to all currently known organisms [18–27]. It is associated with the reading frame of a gene [15]. The reason is both in genetic code structure, which is almost the same for prokaryotes and eukaryotes, and in certain amino acids saturation of proteins [28–30]. We will notice if there is a triplet periodicity phase shift in a gene because it will lead to a shift between the triplet periodicity and the reading frame. This shift may remain for a long time since triplet periodicity of DNA sequence is rather difficult to be changed with a small number of point mutations [31]. The presence of such shift between the triplet periodicity of nucleotide sequence and the reading frame may indicate a frameshift in the analyzed gene [17].

However, the proposed mathematical method which allows us to detect phase shifts of triplet periodicity, is not free from some drawbacks. The main drawback is that the developed method can find a phase shift of a reading frame, created by insertion of rather short sequences (not multiple of three bases), with the length of less than several tens of nucleotides. If there was a long insertion it can significantly change the frequency of the triplet region around the phase shift of the triplet periodicity. So the frameshift detection by this method is complicated.

We set two objectives in this work. First, we wanted to improve the previously developed mathematical method so that it could find potential phase shifts of triplet periodicity [17] to account for possible frameshifts, resulting from rather large insertions of DNA fragments (more that 100 bases of DNA). Second, we wanted to test by the improved algorithm the presence of pair points of phase shift of triplet periodicity, the formation of which may be due to long insertions of DNA fragments in the gene. We analyzed genes from 17 bacterial genomes. Approximately 1% of these genes have pair phase shifts of triplet periodicity, which can be caused by insertion of rather long DNA fragments.

METHOD TO SEARCH FOR PAIR POINTS OF TRIPLET PERIODICITY PHASE SHIFTS IN GENES

We suppose that we have a nucleotide sequence $S = \{s(k), k = 1,2,\ldots l\}$. Each base $s(k)$ is selected from the alphabet $A = \{a,t,c,g\}$, where l is the length of the sequence *S*. We introduce three reading frames in the sequence of *S* and denote them as T_1, T_2 and T_3. Base $s(1)$ of the sequence *S* is the first, second and third codon base for reading frames T_1, T_2 and T_3, respectively. Reading frame T_1 actually exists in the sequence *S*, and the reading frames T_2 and T_3 can be regarded as hypothetical.

We also define three triplet periodicity matrices $M_1(i_1,i_2)$, $M_2(i_1,i_2)$ and $M_3(i_1,i_2)$ for fragment of sequence *S* in the coordinates of i_1 to i_2 in the frames T_1, T_2 and T_3 respectively. We denote this fragment as $S(i_1,i_2)$. Elements of matrices $m_1(i,j)$, $m_2(i,j)$ and $m_3(i,j)$ show the number of bases of type *i* in the sequence *S* ($i=1$ for a, $i=2$ for t, $i=3$ for c, $i=4$ for g), which occurs in codon position *j* (*j* can be 1, 2 or 3) for reading frames T_1, T_2 and T_3, respectively [15]. The phases of triplet periodicity in matrices M_1, M_2, M_3 are coordinates of the first positions in the codons of the sequences $s(1)$, $s(2)$ and $s(3)$. These sequences $s(1)$, $s(2)$ and $s(3)$ are in the reading frames T_1, T_2 and T_3, respectively. Matrices M_1, M_2, M_3 have phases 1, 2 and 3, respectively.

In order to find a pair of points of the phase shift of the triplet periodicity in a symbolic sequences $S(x)$ we suggest a measure of similarity between two triplet periodicity matrices.

1. The measure of similarity between triplet periodicity matrices

To search for a pair of points of the phase shift, we propose a measure of similarity between triplet periodicity matrices. We denote this measure as *U*. If there is no phase shift of the triplet periodicity at *x* then following condition must be satisfied:

$$\begin{cases} U(M_1(x-L,x), M_1(x+1,x+1+L)) \geq U_0 \\ U(M_1(x-L,x), M_2(x+2,x+2+L)) < U_0 \\ U(M_1(x-L,x), M_3(x+3,x+3+L)) < U_0 \end{cases} \quad (1)$$

This condition means that the triplet periodicity matrices to the left and to the right of the point x, defined on the reading frame T_1, are similar to each other. We define a quantitative measure of the similarity between triplet periodicity matrices. Thereto we convert each matrix in (1) into a matrix, where each element will be the argument of the normal distribution. To convert matrix elements we use the approximation of the binomial distribution as follows:

$$n(i,j) = \frac{m(i,j) - Lp(i,j)}{\sqrt{Lp(i,j)(1-p(i,j))}} \quad (2)$$

$$p(i,j) = \frac{x(i)y(j)}{L^2} \quad (3)$$

where $m(i,j)$ is an element of the matrices $M_1(x-L,x)$, $M_1(x+1,x+1+L)$, $M_1(x+2,x+2+L)$ and $M_1(x+3,x+3+L)$; $n(i,j)$ is a normally distributed variable; $x(i,j) = \sum_{i=1}^{4} m(i,j)$, $y(i,j) = \sum_{j=1}^{3} m(i,j)$. For the matrices $M_1(x-L,x)$, $M_1(x+1,x+1+L)$, $M_1(x+2,x+2+L)$ and $M_1(x+3,x+3+L)$ we have the matrices V_1, W_1, W_2 and W_3. Then, for each element of the matrices V_1 and W_k (k = 1, 2, 3) we can obtain:

$$z_{1k}(i,j) = v_1(i,j)w_k(i,j) \quad (4)$$

Probability density function of the product of two independent random variables distributed normally is:

$$f(z) = p^{-1}K_0(|z|) \quad (5)$$

where K_0 is the modified Bessel function of the second kind (MacDonald function). This allows us to calculate the probability that $P(z > z_{1k}(i,j))$. Then we calculate the inverse function of the normal distribution and find the argument of the normal distribution $x_{1k}(i,j)$ for which $P(x > x_{1k}(i,j)) = P(z > z_{1k}(i,j))$. Then we calculated the value:

$$D(1,k) = x_{1k}(i,j) \tag{6}$$

If the hypothesis that the matrices V_1 and W_k (k = 1, 2, 3) are random uncorrelated with each other is true in this case $D(1, k)$ has approximately noncentral normal distribution with mean equal to zero and variance equal to 6.0. Thus, $P(X > D(1,k)), X \sim N(0,6)$ is a probability that similarity of matrices is based on random factors. If $D(1,k)$ is sufficiently big, then the probability that two matrices are similar to each other at random can be rejected. We can take $D(1,k)$ as a measure U in formula (1) from comparing two matrices.

2. Method to search for pair points of triplet periodicity phase shifts in the genes

We select positions $x = 9n + 1$ in sequence S, where $n = 0,1,2,3,\ldots$ For each position in the interval from x to $x + 59$, we make a triplet periodicity matrix M_n. Thus, we make $K = (L-60)/9 + 1$ matrices. Then we compare each matrix with each matrix and calculate three matrices $Sim(1,1)$, $Sim(1,2)$, $Sim(1,3)$ according to formula (6). Second triplet periodicity matrix was taken without a shift, with a shift by one base or with a shift by two bases, respectively. Dimensions of each matrix are (KxK). Matrices $Sim(1,1)$, $Sim(1,2)$, $Sim(1,3)$ show similarity of matrix with index i to matrix with index j (taken without a shift or with a shift by one base or with a shift by two bases, respectively), defined by the formula (6). Then, in the range from 1 to K, we allocate two points of k_1 and k_2, $k_1 \neq k_2$. We calculate the following values:

$$W_1 = \sum_{1\le i<k_1}\sum_{1\le j<k_1} Sim_{ij}(1,1) + \sum_{k_1\le i\le k_2}\sum_{k_1\le j\le k_2} Sim_{ij}(1,1) + \sum_{k_2<i\le K}\sum_{k_2<j\le K} Sim_{ij}(1,1) \tag{7}$$

$$W_{11} = \sum_{1\le i<k_1}\sum_{k_1\le j\le k_2} Sim_{ij}(1,2) + \sum_{1\le i<k_1}\sum_{k_2<j\le K} Sim_{ij}(1,3) + \sum_{k_1\le i\le k_2}\sum_{k_2<j\le K} Sim_{ij}(1,2) \tag{8}$$

$$W_{12} = \sum_{1\le i<k_1}\sum_{k_1\le j\le k_2} Sim_{ij}(1,2) + \sum_{1\le i<k_1}\sum_{k_2<j\le K} Sim_{ij}(1,1) + \sum_{k_1\le i\le k_2}\sum_{k_2<j\le K} Sim_{ij}(1,3) \tag{9}$$

$$W_{21} = \sum_{1\le i<k_1}\sum_{k_1\le j\le k_2} Sim_{ij}(1,3) + \sum_{1\le i<k_1}\sum_{k_2<j\le K} Sim_{ij}(1,1) + \sum_{k_1\le i\le k_2}\sum_{k_2<j\le K} Sim_{ij}(1,2) \tag{10}$$

$$W_{22} = \sum_{1\le i<k_1}\sum_{k_1\le j\le k_2} Sim_{ij}(1,3) + \sum_{1\le i<k_1}\sum_{k_2<j\le K} Sim_{ij}(1,2) + \sum_{k_1\le i\le k_2}\sum_{k_2<j\le K} Sim_{ij}(1,3) \tag{11}$$

$$W_3 = \sum_{1\le i<K}\sum_{1\le j<K} Sim_{ij}(1,1) \tag{12}$$

Then we calculate totals V_{11}, V_{12}, V_{21}, V_{22}.

$$V_{11} = W_1 + W_{11} - W_3 \tag{13}$$
$$V_{12} = W_1 + W_{12} - W_3 \tag{14}$$
$$V_{21} = W_1 + W_{21} - W_3 \tag{15}$$
$$V_{22} = W_1 + W_{22} - W_3 \tag{16}$$

The sum V_{11} corresponds to the situation when there are insertions of one DNA base in the points k_1 and k_2. The sum V_{12} corresponds to the situation when there is an insertion of one DNA base in the point k_1 and an insertion of two DNA bases in the point k_2. The sum V_{21} corresponds to the situation when there is an insertion of two DNA bases in the point k_1 and an insertion of one DNA base in the point k_2. The sum V_{22} corresponds to the situation when there are insertions of two DNA bases in the points k_1 and k_2. Sums V_{11}, V_{12}, V_{21}, V_{22} were reduced by W_3.

We do this summation for all values of k_1 and k_2, and for the sum of V_{11}, V_{12}, V_{21}, V_{22} and find the coordinates $k_1^{\max}$ and $k_2^{\max}$, where these sums have maximum values $V_{11}^{\max}$, $V_{12}^{\max}$, $V_{21}^{\max}$, $V_{22}^{\max}$.

To estimate the statistical significance of variables $V_{11}^{\max}$, $V_{12}^{\max}$, $V_{21}^{\max}$, $V_{22}^{\max}$ we generate random sequences with the same length as the analyzed gene and with the same level of triplet periodicity. We divide gene into three subsequences to create such random sequences. The first of them, denote it as C_1, consists of bases with coordinates $i = 1 + 3n$. The second one – C_2 consists of bases with coordinates $i = 2 + 3n$. And the third C_3 consists of bases with coordinates $i = 3 + 3n$.

Next, we create sequences R_1, R_2 and R_3 by a pseudorandom number generator, which will have the same length as the sequences C_1, C_2, C_3. Then we arrange the sequences R_1, R_2 and R_3 in ascending order, and memorize the permutations made in each sequence. After this, we rearranged the nucleotides in the sequences C_1, C_2 and C_3 as we did when arranging the sequences R_1, R_2 and R_3 in ascending order. After such mixing sequences R_1, R_2 and R_3 we create a random sequence R. There are nucleotides from R_1 at positions $i = 1 + 3n$ in the R sequence. Nucleotides from R_2 are at positions $i = 2 + 3n$ in the R. And nucleotides from R_3 are at positions $i = 3 + 3n$ in the R. The length of the sequence R is equal to the length of the original gene, and nucleotide composition was kept the same as in the original gene.

We generate 500 R sequences for each gene. For each sequence R, we calculate $V_{11}^{\max}$, $V_{12}^{\max}$, $V_{21}^{\max}$, $V_{22}^{\max}$ and then calculate the value:

$$Z_k = \frac{V_k^{\max} - \overline{V_k^{\max}}}{\sqrt{D\left(V_k^{\max}\right)}} \tag{17}$$

Where k takes the values 11, 12, 21 or 22. Values $\overline{V_k^{\max}}$ and $D\left(V_k^{\max}\right)$ for each gene were defined on the set of sequences R.

3. Using Monte Carlo method to determine the threshold value Z_k

To find a threshold values for Z_k, we use genes from 17 bacterial genomes. In total, these genomes contained 66,936 genes. We create random sequences by mixing each gene. It allows to keep the same length distribution of random sequences and the same frequency distribution of bases as in the data bank Kegg. We shuffle sequence in the same manner as described in paragraph 2 to save triplet periodicity in a random sequence. This method allows to save triplet periodicity of the sequence, and to leave only those phase shifts of triplet periodicity, which are associated with random factors. After creating a bank of random sequences, we obtain the levels Z_{11}, Z_{12}, Z_{21}, Z_{22}, for which the number of genes with pair phase shifts is about 18% of what we found for the gene sequences of 17 genes of bacteria. We denote these levels as Z_{11}^0, Z_{12}^0, Z_{21}^0, Z_{22}^0. These levels are 3.1, 2.4, 2.5 and 2.6, respectively. The level of 18% was chosen in order to be able to compare the results in this study with results obtained previously [17].

4. Searching for single phase shifts of triplet periodicity

Single phase shifts of triplet periodicity may cause Z_k value to exceed the designate threshold. Therefore all the found genes with pair shifts of triplet periodicity phase are tested for existence of single shifts of triplet periodicity phase. Searching for single phase shifts is similar to clause 2, but instead of W_k value we use the value of

$$W_k = \sum_{1\le i<k_1}\sum_{1\le j<k_1} Sim_{ij}(1,1) + \sum_{1\le i<k_1}\sum_{k_1\le j\le K} Sim_{ij}(1,k) + \sum_{k_1\le i\le K}\sum_{k_1\le j\le K} Sim_{ij}(1,1) \tag{18}$$

In this case k value is 1 or 2 that corresponds to insertion of one or two DNA bases at the point with position of k. Next we calculate the value of

$$V_k = W_k - W_3 \tag{19}$$

The variable k takes the value 1 and 2 here. Then Z_k value is computed by (16) for these values. Threshold for Z_k (paragraph 3) proves to be 3.8. If the values of $V_{11}^{\max}$, $V_{12}^{\max}$, $V_{21}^{\max}$, $V_{22}^{\max}$ exceed the values of Z_1 и Z_2, we consider that there will be a pair shift of triplet periodicity phase in S sequence; if not, we conclude that there will be a single shift.

5. Drawing $I_{11}(x_1,x_2)$, $I_{12}(x_1,x_2)$ and $I_{13}(x_1,x_2)$ graphs

We built contour graphs of $D(1,k)$, where k = 1,2,3, using (6) to illustrate the found pair frameshifts. These graphs are for S_1 и S_2 subsequences 120 (L_1) units long, which start at x_1 and x_2 in S sequence respectively. In this case S_1 and S_2 are apart and differ at $(x_1, x_1 + L_1 - 1)$ and $(x_2, x_2 + L_1 - 1)$ points. Then $M_1(x_1, x_1 + L_1 - 1)$, $M_1(x_2, x_2 + L_1 - 1)$, $M_2(x_2, x_2 + L_1 - 1)$ and $M_3(x_2, x_2 + L_1 - 1)$ matrices and the value of $D(1,k)$ by (6) are calculated. x_1 и x_2 coordinates change independently from 1 to $L - L_1 + 1$ by 3 bases. It means $x_1 = 1 + 3i$, $i = 0,1,2,3,\ldots$, and $x_2 = 1 + 3j$, $j = 0,1,2,3,\ldots$, where i and j are natural numbers. We have three contour graphs as the result. The first one – $I_{11}(x_1,x_2)$ – shows the similarity between triplet periodicity matrices for T_1 reading frame at different points of S sequence. This graph shows that the segments having phase shifts will have low similarity with the other gene sequences. The second graph $I_{12}(x_1,x_2)$ shows the similarity between triplet periodicity matrices of S sequence calculated by T_1 and T_2 reading frames. And the third graph $I_{13}(x_1,x_2)$ shows the same for T_1 and T_3 reading frames. These contour graphs allow detecting the areas of triplet periodicity phase shifts in S sequence.

RESULTS AND DISCUSSION

This paper presents the results of searching for pair shifts of triplet periodicity phase in bacterial genes. We examined 1 + 1, 1 + 2, 2 + 1 and 2 + 2 shifts. These digits are the numbers of DNA bases that reading frame moves to the right after $k_1^{\max}$ and after $k_2^{\max}$ respectively. Triplet periodicity phase restores in gene after 1 + 2 and 2 + 1 shifts while after 1 + 1 and 2 + 2 shifts it will move one or two DNA bases to the right relative to the phase that is before $k_1^{\max}$ and $k_2^{\max}$ points.

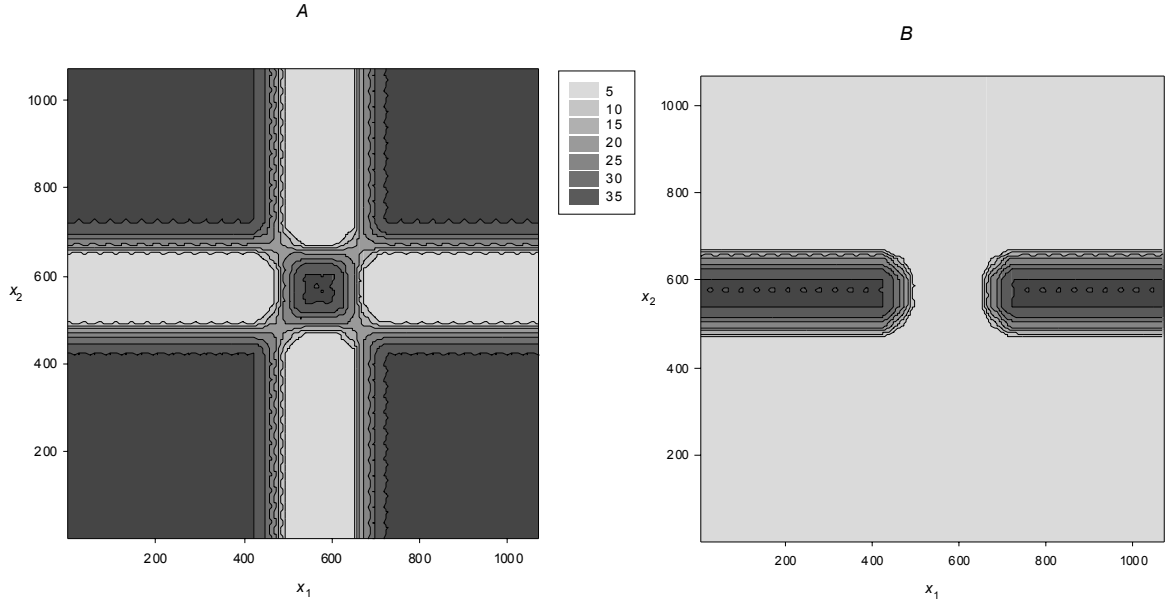

Fig. 1. A. The plot of $I_{11}(x_1,x_2)$ for sequence $(atcgtaggt)_{60}a(atcgtaggt)_{20}aa(atcgtaggt)_{53}$, which has phase shifts after 540 and 721 bases by one and two DNA bases respectively. **B.** The plot of $I_{12}(x_1,x_2)$ for the same sequence. The area between the shifts is dark colored.

First we search for pair phase shifts in artificial periodic sequence. To do this we used our method on $(atcgtaggt)_{134}$ periodic sequence. We haven't found any single or pair shifts of triplet periodicity phase in this case. Then we placed an *a* DNA base into this sequence after the 540th nucleotide, and triplet periodicity phase was shifted by one base, and *aa* DNA bases after the 721th nucleotide that shifted the phase by two bases more and it returned to its initial state. The contour graphs $I_{11}(x_1,x_2)$ and $I_{12}(x_1,x_2)$ for this sequence are on figure 1. As it is shown on figure 1A there won't be similarity between triplet periodicity matrices $M_1(x_1, x_1 + L_1 - 1)$ and $M_1(x_2, x_2 + L_1 - 1)$ if x_1 or x_2 coordinates are inside the area of triplet periodicity phase shift. The graph of $I_{12}(x_1,x_2)$ on figure 1B shows that there will be similarity between matrices $M_1(x_1, x_1 + L_1 - 1)$ и $M_2(x_2, x_2 + L_1 - 1)$ if either x_1 or x_2 is inside the area of triplet periodicity phase shift. There is no graph $I_{13}(x_1,x_2)$ because it is similar to $I_{12}(x_1,x_2)$ and could be made of it by mapping relative to main diagonal.

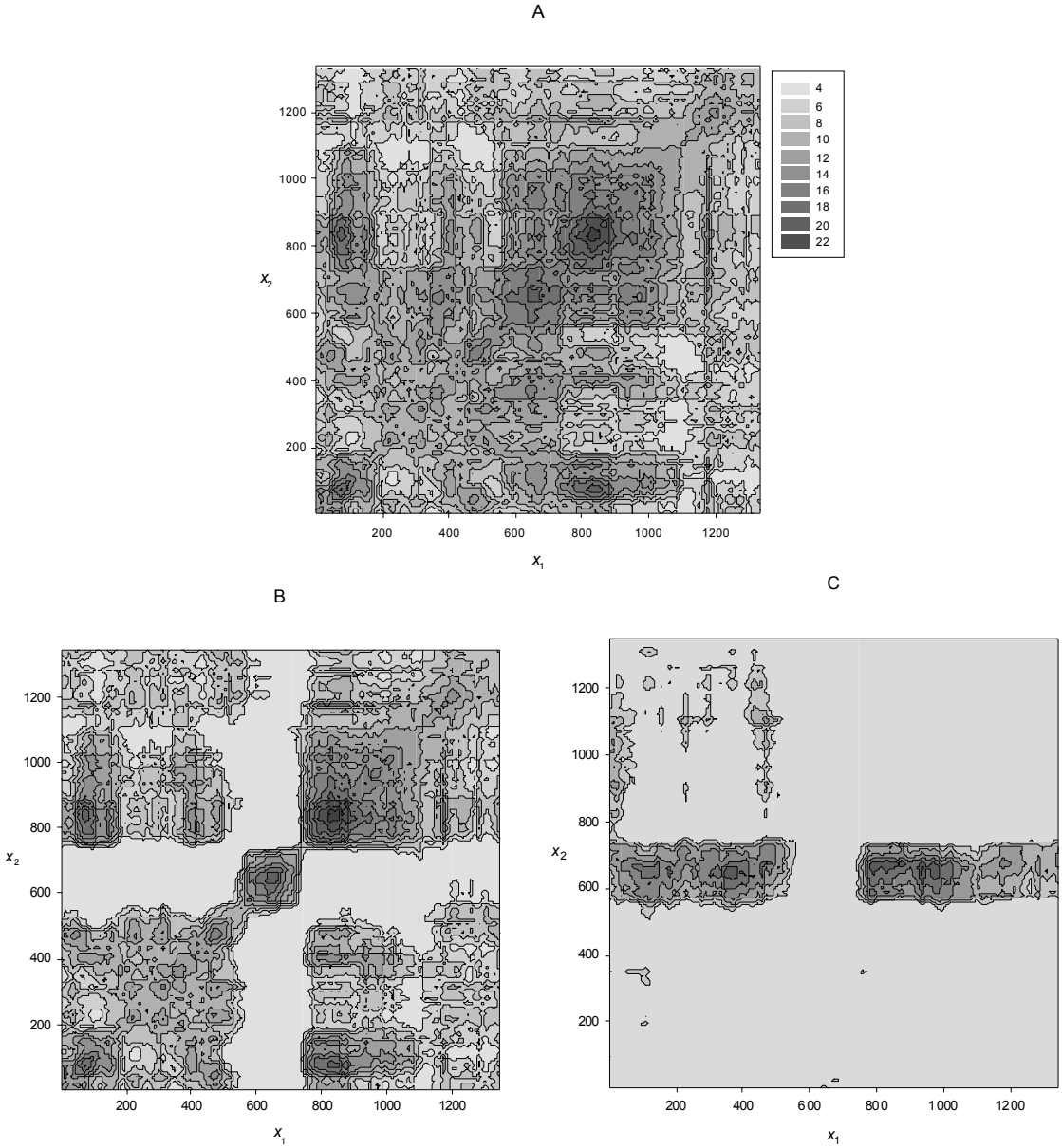

Fig. 2. A. The plot of $I_{11}(x_1,x_2)$ for the sequence of the gene coding chitosanase from *B. subtilis* genome BSU00090 from Kegg data bank. We can see that triplet periodicity is uniform on all the gene length. **B.** The plot of $I_{11}(x_1,x_2)$ for sequence of the gene coding chitosanase from *B. subtilis* genome BSU00090 from Kegg data bank, where we made an insertion of *a* after 600 base and an insertion of *aa* after 781th base. **C.** The plot of $I_{12}(x_1,x_2)$ for sequence of the gene coding chitosanase from *B. subtilis* genome BSU00090 from Kegg data bank, where we made two insertions. The area between the shifts is dark colored.

Next we tried gene sequence from *B. subtilis* genome, BSU26890 from Kegg database, (this gene codes chitosanase). There were no pair shifts of triplet periodicity. We placed an *a* DNA base after 600 nucleotide and *aa* after 781 nucleotide. The contour plot (on figures 2A and 2B) of $I_{11}(x_1,x_2)$ for this sequence shows the situation before and after triplet periodicity phase shift points. There is also $I_{12}(x_1,x_2)$ plot on figure 2C. The area between $k_1^{\max}$ and $k_2^{\max}$ points is marked on figure 2C with the dark horizontal stripe as well as on figure 1B. These examples show that the developed mathematical method can detect phase shifts of triplet periodicity in genes.

After that we analyzed 17 bacterial genomes which contain 66936 genes total. The list of bacteria analyzed in this research is in table 1. We have found 2119 genes containing statistically significant single or pair shifts of triplet periodicity phase. It is about 3 percent of all the examined genes. These data agree with our previous results [17]. We compared the lists of single and pair phase shifts of triplet periodicity we had found before and we obtained in this research. The overlap in these data is about 50%. Such difference may be due to the following reasons. Previously the measure of the difference between triplet periodicity matrices was used to detect phase shifts of triplet periodicity. In this case some shifts could be considered as change points [32]. Also some single shifts could be missed. In this research we measured similarity between matrices. We believe it provides more respective results in searching for single and pair shifts of triplet periodicity phase. Measuring of similarity between matrices made possible to localize the shift coordinate in gene sequence more accurately comparing to previous methods [17].

Table 1. Numbers of pair phase shifts of triplet periodicity in the analyzed bacterial genomes

N	Genome	Amount of genes with single and pair phase shifts of triplet periodicity
1	*A. butzleri*	32
2	*A. vinelandii*	111
3	*B. avium*	88
4	*B. mallei*	204
5	*B. subtilis*	102
6	*E. coli*	149
7	*L. fermentum*	56
8	*M. capsulatus*	117
9	*P. aeruginosa*	83
10	*S. aureus col*	70
11	*S. enterica choleraesuis*	202
12	*S. pneumoniae*	74
13	*S. sonnei*	221
14	*S. typhimurium*	150
15	*V. cholerae*	164
16	*X. campestris*	163
17	*Y. pseudotuberculosis ypiii*	133

There are about 1 percent of genes with pair phase shifts among all the examined genes as it is shown in table 2. Pair shifts of triplet periodicity phase examples are on figures 3 and 4. These figures show that there are DNA parts in gene sequences having triplet periodicity phase shifts. The number of pair shifts is more than it could be expected if triplet periodicity phase shifts are random. Single shifts have been detected in about 2 percent of genes, so we could expect there would be about 0.04 percent of genes having pair phase shifts if shifts are random. The number of pair phase shifts is about 25 times more than the one having been computed by the model where shifts are random. Such difference may be an argument for the

assumption that there are some ferments trying to eliminate frameshifts in genes (trying to repair genes) [33].

Table 2. Numbers of single and pair phase shifts of triplet periodicity in genes from 17 bacterial genomes

Single phase shifts of triplet periodicity	Pair phase shifts of triplet periodicity of the type 1 + 1	Pair phase shifts of triplet periodicity of the type 1 + 2	Pair phase shifts of triplet periodicity of the type 2 + 1	Pair phase shifts of triplet periodicity of the type 2 + 2
1374	185	239	142	179

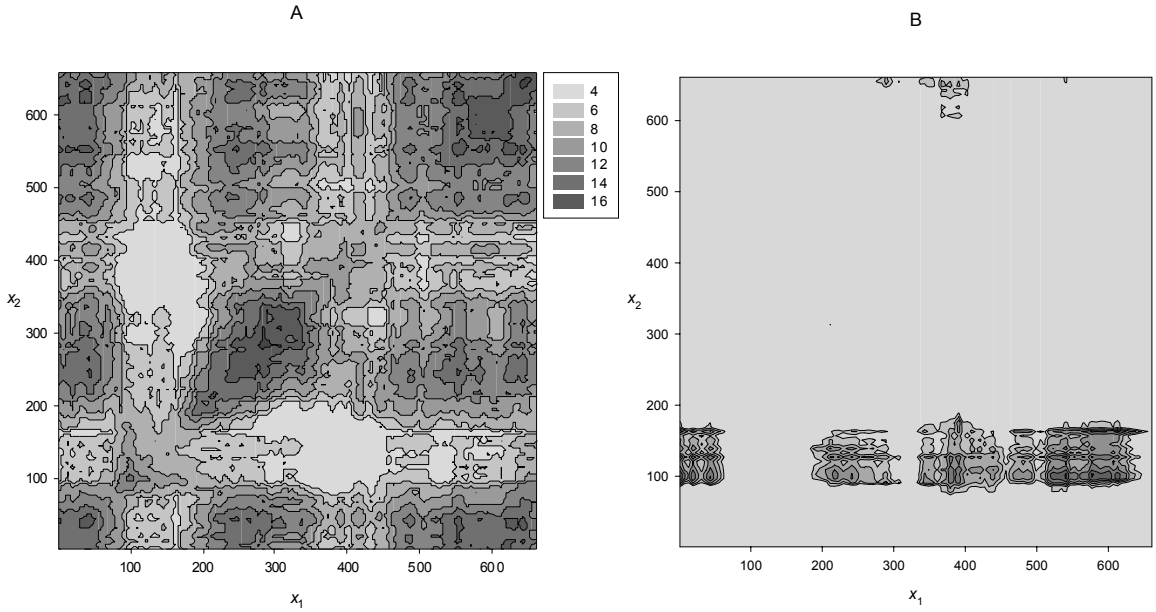

Fig. 3. A. The plot of $I_{11}(x_1,x_2)$ for sequence of the gene coding enoyl-CoA hydratase from *B. avium* genome BAV1386 in Kegg data bank. We can see that this gene contains phase shift of triplet periodicity by one base after ~100th base and by two bases after ~200th base. **B.** The plot of $I_{12}(x_1,x_2)$ for sequence of the gene coding enoyl-CoA hydratase from *B. avium* genome BAV1386 in Kegg data bank. The area between the shifts is dark colored.

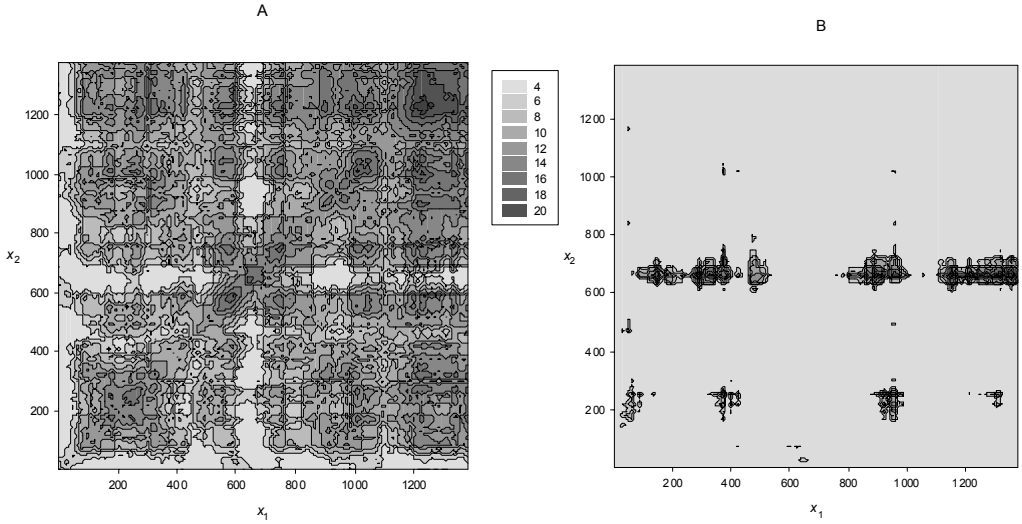

Fig. 4. A. The plot of $I_{11}(x_1,x_2)$ for sequence of the gene coding citrate synthase family protein from *B. mallei* genome BMA2258 in Kegg data bank. We can see that this gene contains a phase shift of triplet periodicity by one base after ~600th base and a shift by two bases after ~700th base. **B.** The plot of $I_{12}(x_1,x_2)$ for sequence of the gene coding citrate synthase family protein from *B. mallei* genome BMA2258 in Kegg data bank. The area between the shifts is dark colored.

t11

We have also examined the similarity of amino acid sequences made for genes having single shifts of triplet periodicity phase for three reading frames. Among the found similarities we picked out only the ones, which started or ended at the point of phase shift (point A). As it shown in table 3 there are about 15 percent of genes with triplet periodicity phase shifts having similarity only with amino acid sequences made for S_1 and S_2 fragments by alternative reading frames. And this similarity ends (for S_1) or starts (for S_2) at the point of phase shift (point A on figure 5A). We have done the same with amino acid sequences made by alternative reading frames for S_1, S_2 and S_3 fragments in the case of pair shifts. These data show that there are amino acid sequences with no pair frameshifts and these sequences are functionally significant.

Table 3. Numbers of alignments for S_1 and S_2 from the genes with triplet periodicity phase shifts. Genes were encoded in three reading frames. S_1 and S_2 were taken as it was described in figure 5A. We used BLAST for searching alignments

	Reading frame		
	1	2	3
S_1	442	29	16
S_2	535	81	135

Table 4. Numbers of alignments for S_1, S_2 and S_3 from the genes with triplet periodicity phase shifts. Genes were encoded in three reading frames. S_1, S_2 and S_3 were taken as it was described in figure 5B. We used BLAST for searching alignments

	Reading frame		
	1	2	3
S_1	168	5	3
S_2	68	9	8
S_3	138	9	12

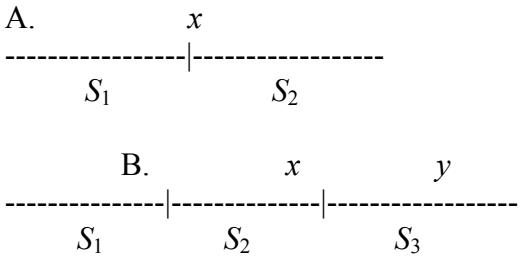

Fig. 5. A. Selection of the parts S_1 and S_2 of the gene if the gene has one phase shift. The first part is before shift point x, the second part is after the point of phase shift of triplet periodicity x. **B.** Selection of the parts S_1, S_2 and S_3 of the gene if the gene has two pair points of phase shift x and y.

We searched with BLAST for alignments of amino acid sequences of genes in which we found the phase shifts of triplet periodicity. Figure 6 shows an example of alignments for S_2 and S_3 fragments of gene coding the inner membrane protein SctQ of the IIIrd type of secretion - BMAA1629 in Kegg database. A pair shift was found in this gene: the first shift is by one nucleotide and the second shift is by two nucleotides. After the first shift the first reading frame near x = 129 nucleotide changes to the second one. After the second shift the reading frame near x = 201 nucleotide changes back to the first one. The alignments for the second and the third fragments were found.

There is also an interesting question about biological significance of pair frameshifts. There are no introns in bacterial genes unlike eukaryote cells so pair frameshifts may be the

way to change a sequence inside a gene. If a gene is not divided by introns it will be difficult to make a significant change in the middle of protein. But pair frameshifts can successfully perform this function and make bacteria adapting better to environmental changes.

```
A.
Query: 137 LPLDAALLPDTVLHLVLSRFDLRLRFETPDAATRHLLCTHGATLHARLDTLL 188
           L LD  +LP+T L L LS   L LRFE     +RHLL  HG TL  R+  LL
Sbjct: 135 LDLDPKILPETRLTLRLSPHTLSLRFEAGHPRSRHLLSEHGDTLRQRIHALL 186

B.
Query: 569 LGEVALPVHVEIDTLSLSIAELAALRPGYVLELPLAARDVPVRLVAYGQAIGGGRLVAVG 628
           L ++ + V  E+    L    L +L PG +++L     D  VRL+A G+ +G GRLV +
Sbjct: 233 LNQLPVQVSFEVGRQILDWHTLTSLEPGSLIDLTTPV-DGEVRLLANGRLLGHGRLVEIQ 291

Query: 629 AHLGVRIDRMA 639
             LGVRI+R+
Sbjct: 292 GRLGVRIERLT 302
```

Fig. 6. A. The alignment between S_2 fragment of gene (137-188 aminoacids) coded in aminoacid sequence in the second frame (reading frame was shifted by one base to the right) and the sequence sp|P35651|HPAP_RALSO HRP-associated protein OS = *Ralstonia solanacearum*. E value = 0.005. **B.** The alignment between S_3 fragment of gene (569-639 aminoacids) coded in aminoacid sequence in the first frame (reading frame was shifted by two bases to the right) and the sequence sp|P40296|YSCQ_YERPS Yop proteins translocation protein Q OS = *Yersinia pseudotuberculosis*. E value = 0.004.

This research results agree with the previous results [17] about the number of genes having frameshifts. As it has been shown before there are 4 percent of genes having at least one frameshift. In this research this number is about 3 percent. Little difference is due to usage different measures of comparing triplet periodicity matrices and different size of analyzed genes. In this research we analyzed genes longer than 60 nucleotides while before we had examined genes longer than 1200 nucleotides only. It is quite logical to think that the probability of frameshift mutations is bigger for longer genes. There is much more significant difference about the quantity of pair frameshifts between this research results and the results of another work [34]. But in this research we used a mathematical measure that made it possible to consider a number of DNA fragment insertions into genes to be pair phase shifts. Just because of this the difference in the data on pair phase shifts is 4-5 times.

We have found only pair phase shifts that are more than 60 DNA bases apart. Also the triplet periodicity must be rather significant to make it possible to find triplet periodicity phase shifts in the gene. At present it is difficult to say if such phase shifts will change the biological function of the encoded protein and if they will create new genes and new biological functions of the encoded proteins. This interesting question can be answered after a series of experiments.

REFERENCES

1. *DNA Repair, Genetic Instability, and Cancer.* Eds. Wei Q., Li L., Chen D.J. World Scientific, 2007.
2. Watson J.D., Levine M., Baker T.A., Gann A., Bell S.P. *Molecular Biology of the Gene.* Benjamin-Cummings Pub. Corp., 2007.
3. Okamura K., Feuk L., Marquès-Bonet T., Navarro A., Scherer S.W. Frequent appearance of novel protein-coding sequences by frameshift translation. *Genomics*. 2006. V. 88. P. 690–697.
4. Raes J., Van de Peer Y. Functional divergence of proteins through frameshift mutations. *Trends Genet*. 2005. V. 21. P. 428–431.

5. Kramer E.M., Su H.-J., Wu C.-C., Hu J.-M. A simplified explanation for the frameshift mutation that created a novel C-terminal motif in the APETALA3 gene lineageBMC. *Evolutionary Biology*. 2006. V. 6 P. 30.
6. States D.J., Botstein D. Molecular sequence accuracy and the analysis of protein coding regions. *Proc. Natl. Acad. Sci., USA*. 1991. V. 88. P. 5518–5522.
7. Pearson W.R., Wood T., Zhang Z., Miller W. Comparison of DNA sequences with protein sequences. *Genomics*. 1997. V. 46. P. 24–36.
8. Birney E., Thompson J., Gibson T. PairWise and SearchWise: finding the optimal alignment in a simultaneous comparison of a protein profile against all DNA translation frames. *Nucl. Acids Res*. 1996. V. 24. P 2730–2739.
9. Guan X., Uberbacher E.C. Alignments of DNA and protein sequences containing frameshift errors. *Comput. Appl. Biosci*. 1996. V. 12. P. 31–40.
10. Antonov I., Borodovsky M. Genetack: frameshift identification in protein-coding sequences by the Viterbi algorithm. *J. Bioinform Comput Biol*. 2010 Jun. V. 8. № 3. P. 535–51.
11. Kislyuk A., Lomsadze A., Lapidus A.L., Borodovsky M. Frameshift detection in prokaryotic genomic sequences. *Int J Bioinform Res Appl*. 2009. V. 5. № 4. P. 458–477.
12. Fichant G.A., Quentin Y. A frameshift error detection algorithm for DNA sequencing projects. *Nucleic Acids Res*. 1995. V. 23. |P. 2900–2908.
13. Médigue C., Rose M., Viari A., Danchin A. Detecting and analyzing DNA sequencing errors: toward a higher quality of the Bacillus subtilis genome sequence. *Genome Res*. 1999. V. 9. P. 1116–1127.
14. Schiex T., Gouzy J., Moisan A., Oliveira Y.D. FrameD: a flexible program for quality check and gene prediction in prokaryotic genomes and noisy matured eukaryotic sequences. *Nucleic Acids Res*. 2003. V. 31. P. 3738–3741.
15. Frenkel F.E., Korotkov E.V. Classification analysis of triplet periodicity in protein-coding regions of genes. *Gene*. 2008. V. 421. P. 52–60.
16. Frenkel F.E., Korotkov E.V. Using triplet periodicity of nucleotide sequences for finding potential reading frame shifts in genes. *DNA Res*. 2009. V. 16. P. 105–114.
17. Korotkov E.V., Korotkova M.A. Study of the triplet periodicity phase shifts in genes. *Journal of Integrative Bioinformatics*. 2010. V. 7. P.131–141.
18. Fickett J.W. Predictive methods using nucleotide sequences. *Methods Biochem Anal*. 1998. V. 39. P. 231–245.
19. Staden R. Statistical and structural analysis of nucleotide sequences. *Methods Mol Biol*. 1994. V. 25. P. 69–77.
20. Baxevanis A.D. Predictive methods using DNA sequences. *Methods Biochem Anal*. 2001. V. 43. P.233–252.
21. Gutiérrez G., Oliver J.L., Marín A. On the origin of the periodicity of three in protein coding DNA sequences. *J Theor Biol*. 1994 Apr 21. V. 167. № 4. P. 413–414.
22. Gao J., Qi Y., Cao Y., Tung W.-W. Protein Coding Sequence Identification by Simultaneously Characterizing the Periodic and Random Features of DNA Sequences. *Journal of Biomedicine and Biotechnology*. 2005. V. 2. P. 139–146.
23. Yin C., Yau S.S. Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence. *Journal of Theoretical Biology*. 2007. V. 247. P. 687–694.
24. Eskesen S.T., Eskesen F.N., Kinghorn B., Ruvinsky A. Periodicity of DNA in exons. *BMC Molecular Biology*. 2004. V. 5. P. 12.
25. Bibb M.J., Findlay P.R., Johnson M.W. The relationship between base composition and codon usage in bacterial genes and its use for the simple and reliable identification of protein-coding sequences. *Gene*. 1984 Oct. V. 30. № 1-3. P. 157–166.
26. Konopka A.K. *S*equences and codes: fundamentals of biomolecular cryptology. In: *Biocomputing: Informatics and genome projects*. Ed. Smith D. San Diego, CA: Academic Press, 1994. P. 119–174.

27. Trifonov E.N. Elucidating sequence codes: three codes for evolution. *Ann NY Acad Sci*. 1999. V. 870. P. 330–338.
28. Eigen M., Winkler-Oswatitsch R. Transfer-RNA: the early adaptor. *Naturwissenschaften*. 1981. V. 68. P. 217–228.
29. Zoltowski M. Is DNA Code Periodicity Only Due to CUF - Codons Usage Frequency? In: *Conf Proc IEEE Eng Med Biol Soc*. 2007. V. 1. P. 1383–1386.
30. Antezana M.A., Kreitman M. The nonrandom location of synonymous codons suggests that reading frame-independent forces have patterned codon preferences. *J Mol Evol*. 1999. V. 49. № 1. P. 36–43.
31. Korotkov E.V., Korotkova M.A., Frenkel F.E., Kudryashov N.A. The Informational Concept of Searching for Periodicity in Symbol Sequences. *Molecular Biology*. 2003. V. 37. P. 372–386.
32. Suvorova Y.M., Rudenko V.M., Korotkov E.V. Detection change points of triplet periodicity of gene. *Gene*. 2012. V. 491. P. 58–64.
33. Strauss B.S. Frameshift mutation, microsatellites and mismatch repair. *Mutation Research*. 1999. V. 437. P. 195–203.
34. Korotkova M.A., Kudryashov N.A., Korotkov E.V. An approach for searching insertions in bacterial genes leading to the phase shift of triplet periodicity. *Genomics Proteomics Bioinformatics*. 2011. V. 9. P. 158–170.

Received May 11, 2012.
Published August 08, 2012.